Time-Dependent Stick-Breaking Processes

J.E. Griffin*

Institute of Mathematics, Statistics and Actuarial Science,

University of Kent, U.K.

M. F. J. Steel

Department of Statistics, University of Warwick, U.K.

Abstract

This paper considers the problem of defining a time-dependent nonparametric prior. A recursive construction allows the definition of priors whose marginals have a general stick-breaking form. The processes with Poisson-Dirichlet and Dirichlet process marginals have interesting interpretations that are further investigated. We develop a general conditional Markov Chain Monte Carlo (MCMC) method for inference in the wide subclass of these models where the parameters of the stick-breaking process form increasing sequences. We derive a Pólya urn scheme type representation of the Dirichlet process construction, which allows us to develop a marginal MCMC method for this case. The results section shows the relative performance of the two MCMC schemes for the Dirichlet process case and contains three real data examples.

*Corresponding author: Jim Griffin, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K. Tel.: +44-1227-823627; Fax: +44-1227-827932; Email: J.E.Griffin-28@kent.ac.uk. The authors would like to acknowledge helpful comments from seminar audiences at the Universities of Newcastle, Nottingham, Bath, Sheffield and Bristol, Imperial College London and the Gatsby Computational Neuroscience group

Keywords: Bayesian Nonparametrics, Dirichlet Process, Poisson-Dirichlet Process, Time-Dependent Nonparametrics.

1 Introduction

Nonparametric estimation is an increasingly important element in the modern statistician's toolbox. The availability of efficient methods for the estimation of unknown distribution has lead to the development of regression methods that allow many unknown distributions to be estimated jointly. Recently, there has been interest in extending standard Bayesian nonparametric methods, particularly the Mixture of Dirichlet Processes (Antoniak 1974) to regression and time series contexts (*e.g.* De Iorio *et al* 2004, Müller *et al* 2004, Griffin and Steel 2006 and Dunson *et al* 2007). The development of methodology is generally difficult because we must define a measure-valued stochastic process. In this paper we will work exclusively on the time series problem, which often allows simpler constructions.

Time-dependent nonparametric priors have been used in a wide-range of applications. Two extreme examples would be: the observation of a single time series over a long time period such as a stock index considered by Griffin and Steel (2006) and Caron *et al* (2007). On the other hand we can consider many observations at each time point *e.g.* travel claims considered by Rodriguez and ter Horst (2008). We assume observations are independently drawn conditional on the unknown distributions. Therefore longitudinal data cannot be directly modelled but can be modelled hierarchically (Dunson 2006) using these models. If we have a long time series, we need strong prior assumptions for effective inference. In the second case, we have more information at each time point (*e.g.* we could make inference about the distribution at each time point independently) but our aim is to estimate and exploit the relationship between distributions at each time point.

Suppose we observe a sample of observations $y_1, \ldots, y_n$ taken at times $t_1, \ldots, t_n$ respectively,

then a Bayesian nonparametric analysis could assume that their conditional can be expressed as

$$f_{t_i}(y_i) = \int k(y_i|\varphi_i) dG_{t_i}(\varphi_i) \tag{1}$$
$$G_{t_i} = \sum_{j=1}^{\infty} p_j(t_i)\delta_{\theta_j(t_i)}$$

for some density function $k(\cdot)$, where $p_1, p_2, p_3, \ldots$ sum to one and are independent of $\theta_1, \theta_2, \theta_3, \ldots$ while δ_θ is the Dirac delta function which places measure 1 on the point θ. This implies that (1) defines a mixture of parametric distributions with mixing distribution G_{t_i}. The model reduces to a species sampling mixture model if all the observations are taken at a single time. The class of species sampling mixture models includes Dirichlet Process Mixtures (Lo 1984), Stick-Breaking Mixtures (Ishwaran and James 2001, 2003), and Normalized Random Measure mixtures (James *et al* 2005). If we assume that $p_j(t) = p_j$ for all t, the model reduces to an infinite mixture of time-series model in the framework of Dependent Dirichlet Processes. Rodriguez and ter Horst (2008) develop methods in this direction. Alternatively, several constructions make the assumption that $\theta_j(t)$ does not depend on time and so $\theta_j(t) = \theta_j$ for all t. Examples in continuous time include the Ornstein-Uhlenbeck Dirichlet Process (Griffin 2007) and the "arrivals" construction of the π-DDP (Griffin and Steel 2006) and in discrete time include the Time Series Dependent Dirichlet Process (Nieto-Barajas *et al*, 2008). Zhu *et al* (2005) model the distribution function of the observables directly by defining a Time-Sensitive Dirichlet Process which generalizes the Pólya urn scheme of the Dirichlet process to

$$G_t|y \propto \mathrm{I}(t_i < t)\exp\{-(t-t_i)\}\delta_{y_i} + MH$$

where t_i is the time at which y_i is observed and M and H are the mass parameter and the centring distribution of the Dirichlet process. The influence of an observation at time t_i decreases as $t - t_i$ becomes large. However, the process is not consistent under marginalisation of the sample. A related approach is described by Caron *et al* (2007). In discrete time, Dunson (2006) proposes the evolution

$$G_t = \pi G_{t-1} + (1-\pi)\epsilon_t$$

where π is a parameter and ϵ_t is a realisation from a Dirichlet process. This defines an AR-process type model and an explicit Pólya urn-type representation allows efficient inference. This has some

similarities to the "arrivals" π-DDP in discrete time, where the model is generalized to

$$G_t = \pi_{t-1} G_{t-1} + (1 - \pi_{t-1})\epsilon_{t-1}$$

where ϵ_t is a discrete distribution with a Poisson-distributed number of atoms and π_t is a random variable correlated with ϵ_t. Griffin and Steel (2006) show how to ensure that the marginal law of G_t follows a Dirichlet process for all t.

This paper extends the literature on time-dependent nonparametric priors in a number of ways: 1) we describe and study in detail a construction of a measure-valued stochastic process with any given stick-breaking process as the marginal process, 2) we define a Markov chain Monte Carlo (MCMC) scheme for inference with a wide subclass of these process (including the Poisson-Dirichlet process), and 3) we derive Pólya urn-type schemes for the important special case of a Dirichlet process marginal.

The paper is organised in the following way: Section 2 describes the link between stick-breaking processes and time-varying nonparametric models and presents a method for constructing stationary processes with a given marginal stick-breaking process. Section 3 discusses two important special cases: Dirichlet process and Poisson-Dirichlet marginals. Section 4 briefly describes the proposed computational methods. In Section 5 we analyse some simulated data and explore three econometric examples using the two leading processes. Proofs of all Theorems are grouped in Appendix A, whereas Appendix B presents details of the MCMC algorithms.

2 Time-dependent nonparametric priors and stick-breaking

In this paper, we will work exclusively with the model in (1) where $\theta_j(t) = \theta_j$ for all t. An important characteristic of (1) is that G_t is discrete at all times, which can make seemingly natural generalisations of standard time series model have unusual properties. For example, we can define a random walk $y_1, y_2, y_3, \ldots$ with normal increments equivalently as:

$$y_i \sim \mathrm{N}(y_{i-1}, \sigma^2), \tag{2}$$

where $\mathrm{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, or

$$y_i = y_{i-1} + \epsilon_i \text{ where } \epsilon_1, \epsilon_2, \ldots \text{ are i.i.d. } \mathrm{N}(0, \sigma^2). \tag{3}$$

This equivalence does not extend naturally to infinite discrete distributions. Suppose that we wanted to generalize (2) we might define

$$G_t | G_{t-1} \sim \mathrm{DP}(M, G_{t-1}) \tag{4}$$

where $\mathrm{DP}(\alpha, G)$ represents a Dirichlet process with mass parameter α and centring distribution G. Under this model, the variance of the mass assigned to some set B will be

$$\mathrm{Var}[G_{t+j}(B)|G_{t-1}] = \alpha_j G_{t-1}(B)\left[1 - G_{t-1}(B)\right],$$

where α_j is an increasing sequence depending on M and tending to 1 as $j \to \infty$. The variability is increasing with j, which we would want, but, since G_{t+j} is a discrete distributions, the maximum variability arises when our distribution is a single point mass at some value drawn from G_{t-1}. A natural generalisation of (3) is

$$G_t = (1 - V_t)G_{t-1} + V_t \epsilon_t$$

where ϵ_t is a random probability distribution (with infinite or finite numbers of atoms) which is independent of G_{t-1}. The latter construction does not suffer from this degeneracy problem. The degeneracy of the process defined by (4) is similar to the phenomenon of particle depletion in particle filters that is caused by the approximation of a continuous distribution by a finite, discrete distribution (see *e.g.* Doucet *et al* 2001). These problems are addressed through a so-called rejuvenation step where randomness is introduced into the distribution of particles. Adding an innovation at each time step will have a similar effect in these models.

We now focus on the choice of V_t and ϵ_t. If ϵ_t is a discrete distribution then G_t will also be discrete at all times. If we further restrict ϵ_t to be a single atom at a point θ_t drawn at random from H then each marginal distribution naturally has a so-called "stick-breaking" structure since we can write

$$G_t = \sum_{i=-\infty}^{t} \delta_{\theta_i} V_i \prod_{j=i+1}^{t} (1 - V_j) \tag{5}$$

where stick-breaking is applied backwards in time (the atom introduced at time t has the largest expected weight). More formally, the stick-breaking construction of random probability measures (Pitman 1996, Ishwaran and James 2001) is defined as follows

Definition 1 *Suppose that* $\mathbf{a} = (a_1, a_2, \dots)$ *and* $\mathbf{b} = (b_1, b_2, \dots)$ *are sequences of positive real numbers. A random probability measure* G *follows a stick-breaking process if*

$$G \stackrel{d}{=} \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$$

$$p_i = V_i \prod_{j<i} (1 - V_j)$$

where $\theta_1, \theta_2, \theta_3, \cdots \overset{i.i.d.}{\sim} H$ *and* $V_1, V_2, V_3, \dots$ *is a sequence of independent random variables for which* $V_i \sim Be(a_i, b_i)$ *for all* i*. This process will be written as* $G \sim \Pi(\mathbf{a}, \mathbf{b}, H)$.

Ishwaran and James (2001) shows that the process is well-defined (*i.e.* $\sum_{i=1}^{\infty} p_i = 1$ almost surely) if

$$\sum_{i=1}^{\infty} \log(1 + a_i/b_i) = \infty.$$

A number of standard processes fall within this class. The two most widely used are the Dirichlet process where $a_j = 1$ and $b_j = M$ and the Poisson-Dirichlet (or Pitman-Yor) process $a_j = 1 - a$, $b_j = M + aj$ (with $0 \leq a < 1, M > -a$). The Dirichlet process has been almost ubiqitous in applications of Bayesian nonparametric model. The Poisson-Dirichlet process was discussed in some detail in Pitman and Yor (1997) and James (2008) and has been used by *e.g.* Teh (2006).

Griffin and Steel (2006) shows that if we choose $V_j \sim \text{Be}(1, M)$ then the process in (5) is stationary in the sense that G_t will have the same stick-breaking prior as G_s for all s and t and "mean-reverting" in the sense that $\text{E}[G_t] \to H$ as $t \to \infty$. They further subordinate the process in (5) to some additive process, X_t, which allows modelling of the autocorrelation structure whilst maintaining a marginal Dirichlet process. If the additive process is a stationary Poisson process with intensity λ then the autocorrelation has an exponential form

$$\text{Corr}(G_t, G_{t+k}) = \exp\left\{-\frac{\lambda k}{M+1}\right\} \equiv \rho^k.$$

This naturally leads to two questions: 1) how can we construct processes whose margins follows a specific stick-breaking process? And 2) how can we control the autocorrelation of the process of distributions $\{G_t\}_{t=-\infty}^{\infty}$?

To develop answers to these questions we will initially work with an extension of the discrete time process in (5) and then extend to continuous time by subordination. The model in (5) naturally includes the marginal two-parameter Beta processes (Ishwaran and Zarepour 2000) which arise by assuming that $V_i \sim \text{Be}(a, M)$. However, if the distributions of the V_i's depends on the ordering, we need to account for that since the ordering of the atoms evolves over time. Thus, such marginal stick-breaking processes can only arise from the generalized model

$$G_t = \sum_{i=-\infty}^{t} \delta_{\theta_i} V_{i,t-i+1} \prod_{j=i+1}^{t} (1 - V_{j,t-j+1}), \tag{6}$$

where $V_{j,1}, V_{j,2}, \ldots$ denotes the value of the break introduced at time j as it evolves over time and $V_{j,t}$ represents its value at time $j + t$. The stochastic processes $\{V_{i,t}\}_{t=1}^{\infty}$ and $\{V_{j,t}\}_{t=1}^{\infty}$ must be independent for $i \neq j$ if the process G_t is to have a stick-breaking form. If we want G_t to follow a stick-breaking process with parameters $\mathbf{a}$ and $\mathbf{b}$ then the marginal distribution of $V_{i,t}$ would have to be distributed $\text{Be}(a_t, b_t)$ for all i and t and clearly we need time dependence of the breaks. The following theorem allows us to define a reversible stochastic process $V_{i,t}$ which has the correct distributions at all time points to define a stationary process whose marginal process has a given stick-breaking form. In the following theorem we define that if X is distributed $\text{Be}(a, 0)$ then $X = 1$ with probability 1 and if X is distributed $\text{Be}(0, b)$ then $X = 0$ with probability 1. We will use the notation $\text{Ga}(a)$ to denote a Gamma distribution with shape a and unitary scale.

Theorem 1 *Suppose that $V_{j,t} \sim Be(a_t, b_t)$ and $V_{j,t+1} \sim Be(a_{t+1}, b_{t+1})$ then the following relationships maintain the marginal distributions:*

Case (i) If $a_{t+1} \geq a_t$ and $b_{t+1} \geq b_t$

$$V_{j,t+1} = w_{j,t+1} V_{j,t} + (1 - w_{j,t+1}) \epsilon_{j,t+1}$$

where $w_{j,t+1} \sim Be(a_t + b_t, a_{t+1} + b_{t+1} - a_t - b_t)$ and $\epsilon_{j,t+1} \sim Be(a_{t+1} - a_t, b_{t+1} - b_t)$.

Case (ii) *If* $a_{t+1} \leq a_t$ *and* $b_{t+1} \leq b_t$

$$V_{j,t+1} = \frac{w_{j,t+1} V_{j,t}}{w_{j,t+1} V_{j,t} + (1 - w_{j,t+1})(1 - V_{j,t})}$$

where $w_{j,t} = \frac{x_{j,t}}{x_{j,t}+z_{j,t}}$ *and* $x_{j,t} \sim$ *Be*$(a_{t+1}, a_t - a_{t+1})$ *and* $z_{j,t} \sim$ *Be*$(b_{t+1}, b_t - b_{t+1})$.

Case (iii) *If* $a_{t+1} \leq a_t$ *and* $b_{t+1} \geq b_t$

$$V_{j,t+1} = \frac{x_{j,t} V_{j,t}}{x_{j,t} V_{j,t} + 1 - x_{j,t} + z_{j,t}}$$

where $x_{j,t} \sim$ *Be*$(a_{t+1}, a_t - a_{t+1})$ *and* $z_{j,t} \sim$ *Ga*$(b_{t+1} - b_t)$.

Case (iv) *If* $a_{t+1} \geq a_t$ *and* $b_{t+1} \leq b_t$

$$V_{j,t+1} = \frac{z_{j,t} V_{j,t}}{x_{j,t}(1 - V_{j,t}) + z_{j,t} + V_{j,t}}$$

where $x_{j,t} \sim$ *Be*$(a_{t+1}, a_t - a_{t+1})$ *and* $z_{j,t} \sim$ *Ga*$(b_{t+1} - b_t)$.

The application of this theorem allows us to construct stochastic processes with the correct margins for any stick-breaking process. We will concentrate on the case where $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ form nondecreasing sequences (case (i)). This is the most computationally convenient since the transitions of the stochastic process are mixtures, for which we have well-developed simulation techniques. The stochastic process can then be explicitly written as

$$V_{j,1} = \epsilon_{j,1}$$

with $\epsilon_{j,1} \sim \text{Be}(a_1, b_1)$ and, for $m > 1$,

$$\begin{aligned} V_{j,m} &= w_{j,m} V_{j,m-1} + (1 - w_{j,m}) \epsilon_{j,m} \\ &= \epsilon_{j,1} \prod_{i=2}^{m} w_{j,i} + \sum_{l=2}^{m} \epsilon_{j,l} (1 - w_{j,l}) \prod_{i=l+1}^{m} w_{j,i}, \end{aligned}$$

where $w_{j,m} \sim \text{Be}(a_{m-1} + b_{m-1}, a_m + b_m - a_{m-1} - b_{m-1})$ and $\epsilon_{j,m} \sim \text{Be}(a_m - a_{m-1}, b_m - b_{m-1})$. Applying Theorem 1 (case (i)) the marginal distribution of $V_{j,m} \sim \text{Be}(a_m, b_m)$. It is interesting to note that $V_{j,m}$ is also a draw from a finite stick-breaking distribution where the atoms are located at $\epsilon_{j,m}, \epsilon_{j,m-1}, \ldots, \epsilon_{j,1}$ with breaks $(1 - w_{j,m}), (1 - w_{j,m-1}), \ldots, (1 - w_{j,2}), 1$. This will be useful for our computational approach using a Markov chain Monte Carlo sampler.

In the same way as Griffin and Steel (2006), we will define a continuous-time process by subordinating this discrete-time process to a Poisson process (other additive processes could be used).

Definition 2 *Assume that $\tau_1, \tau_2, \tau_3, \ldots$ follow a homogeneous Poisson process with intensity λ and the count $m_j(t)$ is the number of arrivals of the point process between τ_j and t, so that*

$$m_j(t) = \#\{k|\tau_j < \tau_k < t\}.$$

Let Π be a stick-breaking process with parameters $\mathbf{a}, \mathbf{b}, H$ where $\mathbf{a}$ and $\mathbf{b}$ are nondecreasing sequences. Define

$$G_t = \sum_{j=1}^{\infty} p_j(t)\delta_{\theta_j}$$

where $\theta_1, \theta_2, \theta_3, \cdots \overset{i.i.d.}{\sim} H$ and

$$p_j(t) = \begin{cases} 0 & \tau_j > t \\ V_j(t)\prod_{k|\tau_j<\tau_k<t}(1 - V_k(t)) & \tau_j < t \end{cases}$$

$$V_j(t) = \sum_{l=1}^{m_j(t)+1} \epsilon_{j,l}(1 - w_{j,l}) \prod_{i=l+1}^{m_j(t)+1} w_{j,i}$$

with $\epsilon_{j,1} \sim Be(a_1, b_1)$, $w_{j,1} = 0$ and $w_{j,m} \sim Be(a_{m-1} + b_{m-1}, a_m + b_m - a_{m-1} - b_{m-1})$, $\epsilon_{j,m} \sim Be(a_m - a_{m-1}, b_m - b_{m-1})$ for $m \geq 2$. Then we call G_t a Π-AR (for Π-autoregressive) process, denoted as Π-AR$(\mathbf{a}, \mathbf{b}, H; \lambda)$.

In the following section we will consider the Dirichlet Process AR (DPAR) and the Poisson-Dirichlet AR (PDAR). The properties of the discrete-time process (*i.e.* stationarity and mean-reversion) will also be true for these subordinated versions. Defining a Π-AR process as the non-parametric component in the hierarchical model in (1) defines a time-dependent version of stick-breaking mixture models. This hierarchical model defines a standard change-point model (*e.g.* Carlin *et al* 1992) if G_t is a single atom for all t. Therefore the new hierachical model defines a generalisation of change-point models that allow observations to be drawn from a mixture distribution where components will change according to a point process. Alternatively, as $\lambda \to 0$, the process tends to the corresponding static nonparametric model and as $\lambda \to \infty$ the process becomes uncorrelated in time. We have restricted attention to increasing sequences since efficient inferential methods can be developed. This definition could be extended to define priors where the marginal process is stick-breaking but the parameters are not increasing using Theorem 1.

3 Special cases

3.1 Dirichlet process marginals

The Dirichlet process is an important special case because it is the most commonly used nonparametric prior for the mixing distribution and our time series model has a simple form in this case. It arises when $a_j = 1$ and $b_j = M$ for all j. The DPAR model then has $\epsilon_{j,0} \sim \text{Be}(1, M)$ and $w_{j,m} = 1$, $\epsilon_{j,m} = 0$ for all $m \geq 1$ so that $V_j(t) = \epsilon_{j,0}$ for all t. Writing $V_j = \epsilon_{j,0}$ motivates the following definition.

Definition 3 *Let $\tau_1, \tau_2, \ldots$ follow a homogeneous Poisson process with intensity λ and $V_1, V_2, \cdots \overset{i.i.d.}{\sim}$ Be$(1, M)$. Then we say that $\{G_t\}_{t=-\infty}^{\infty}$ follows a DPAR$(M, H; \lambda)$ if*

$$G_t = \sum_{j=1}^{\infty} p_j(t)\delta_{\theta_j}$$

where $\theta_1, \theta_2, \cdots \overset{i.i.d.}{\sim} H$ and

$$p_j(t) = \begin{cases} 0 & \tau_j > t \\ V_j \prod_{\{k|\tau_j < \tau_k < t\}} (1 - V_k) & \tau_j < t. \end{cases}$$

An important property of the Dirichlet process is the availability of a Pólya urn scheme representation, which integrates out the unknown random measure to describe the probability law of a sequence of draws from the unknown random measure. Importantly, this defines a finite-dimensional representation of the process. We can also describe the DPAR model with a finite-dimensional representation of a sequence of draws from the process.

Let $\zeta_1, \zeta_2, \ldots, \zeta_n$ be the sequence of draws where ζ_i is drawn at time t_i so that $\zeta_i \sim G_{t_i}$. The process is more easily expressed by working with allocations $s_1, s_2, \ldots, s_n$ which are chosen so that $\zeta_i = \theta_{s_i}$. Let $\mathcal{T} = \{\tau_1, \tau_2, \ldots\}$ and let $\mathcal{T}_n = \{\tau_{s_1}, \tau_{s_2}, \ldots, \tau_{s_n}\}$ which is the set of all times which have observations allocated to them. The size of $\mathcal{T}_n$ is k_n. We also define an active set,

$$A_n(t) = \{i | t_i \geq t \text{ and } \tau_{s_i} < t \text{ for } 1 \leq i \leq n\},$$

which contains the observations available to be allocated at time t. If all observations are made at the same time, the model reduces to the standard Dirichlet process and $A_n(t)$ will be an increasing

process. However, once we make observations at different times then the process will still be increasing between observed times but it will be decreasing at these times. It will be useful to find the posterior distribution of the point process $\mathcal{T}$ given the first n allocations. Let $\mathcal{S}_{n,m} = \mathcal{T}_n \cup \{t_1, \ldots, t_m\}$ which has size $l_{n,m}$, and let $\mathcal{T}_n^C$ be the set difference of $\mathcal{T}$ and $\mathcal{T}_n$.

Theorem 2 *Let $\phi_1 < \phi_2 < \cdots < \phi_{l_{n,m}}$ be the elements of $\mathcal{S}_{n,m}$. Then $\mathcal{T}_n^C$ conditional on $s_1, s_2, \ldots, s_n, \mathcal{S}_{n,m}$ is an inhomogeneous Poisson process with a piecewise constant intensity, $f(x)$, where*

$$f(x) = \begin{cases} \lambda & -\infty < x \leq \phi_1 \\ \lambda \left(\frac{M}{M + A_n(\phi_i)} \right) & \phi_{i-1} < x \leq \phi_i, \quad 2 \leq i \leq l_{n,m} \\ \lambda & \phi_{l_{n,m}} < x < \infty. \end{cases}$$

This theorem implies that the intensity of the underlying Poisson process falls as $A_n(\phi_i)$ increases. The size of the decrease is also influenced by M, with larger values of M leading to smaller decreases. The predictive distribution of s_{n+1} conditional on $s_1, s_2, \ldots, s_n$ and the set of times $\mathcal{S}_{n,m}$ can be represented as a finite dimensional distribution. The new observation t_{m+1} can be allocated to points in $\mathcal{T}_n$ or to a new point which is drawn from $\mathcal{T}_n^C$. In this way, we provide a generalized Pólya urn scheme.

Theorem 3 *Let $\tau_1^\star < \tau_2^\star < \cdots < \tau_{k_n}^\star$ be the elements of $\mathcal{T}_n$ and $\phi_1 < \phi_2 < \cdots < \phi_{l_{n,m+1}}$ be the elements of $\mathcal{S}_{n,m+1}$. We define*

$$\begin{aligned} C_i &= \exp\left\{ -\frac{\lambda M^2 (\phi_{i+1} - \phi_i)}{(M + A_n(\phi_{i+1}))^2 (1 + M + A_n(\phi_{i+1}))} \right\} \\ &= \rho^{\frac{M^2(M+1)}{(M+A_n(\phi_{i+1}))^2(1+M+A_n(\phi_{i+1}))}(\phi_{i+1} - \phi_i)}, \qquad 1 \leq i < l_{n,m+1} \end{aligned}$$

and

$$D_i = \frac{M + A_n(\tau_i^\star)}{1 + \eta_i + M + A_n(\tau_i^\star)}, \qquad 1 \leq i \leq k_n$$

where $\eta_i = \sum_{j=1}^n I(s_j = i)$. Let $\phi_p = t_{m+1}$ and $\tau_q^\star$ be the largest element of $\mathcal{T}_n$ smaller than t_{m+1}. The predictive distribution can be represented by

$$p(s_{n+1} = j) = (1 - D_j) \prod_{\{h | \tau_j^\star \leq \phi_h \leq \phi_p\}} C_h \prod_{i=j+1}^{q} D_i, \qquad 1 \leq j \leq q.$$

$$p(s_{n+1} = k_n + 1 \text{ and } \tau^\star_{k_n+1} \in (\phi_j, \phi_{j+1})) = (1 - C_j) \prod_{h=j+1}^{p} C_h \prod_{\{i|\phi_j < \tau^\star_i \leq \phi_p\}} D_i$$

$$p\left(s_{n+1} = k_n + 1 \text{ and } \tau^\star_{k_n+1} \in (-\infty, \phi_1)\right) = \prod_{h=1}^{p} C_h \prod_{i=1}^{q} D_i.$$

Let $TEx_{(a,b)}(\lambda)$ *represent an exponential distribution truncated to* (a, b) *with p.d.f.*

$$f(x) \propto \lambda \exp\{-\lambda x\}, \qquad a < x < b.$$

If $s_{n+1} = k_n + 1$ *then the new time* $\tau^\star_{k_n+1}$ *is drawn as follows: If* $\tau^\star_{k_n+1} \in (-\infty, \phi_1)$, $\tau^\star_{k_n+1} = \phi_1 - x$ *where* $x \sim Ex\left(\frac{\lambda}{M+1}\right)$, *which represents an exponential distribution with mean* $(M + 1)/\lambda$, *and if* $\tau^\star_{k_n+1} \in (\phi_j, \phi_{j+1})$, $\tau^\star_{k_n+1} = \phi_{j+1} - x$ *where* $x \sim TEx_{(0,\phi_{j+1}-\phi_j)}\left(\frac{\lambda M^2}{(M+A_n(\phi_{j+1}))^2(M+A_n(\phi_{j+1})+1)}\right)$.

This representation of the predictive distribution has a stick-breaking structure with a finite number of breaks. The representation defines a probability distribution for any choice of C and D. However, it is not clear whether other choices of C and D would follow from a time series model whose marginals define exchangeable sequences drawn from random probability measures.

3.2 Poisson-Dirichlet marginals

The Poisson-Dirichlet process was introduced into the Bayesian nonparametric literature by Ishwaran and James (2001) as an alternative prior to the Dirichlet process for the mixing distribution in a mixture model. A particular feature of the process is the sub-geometric rate of decays of the expected probabilities defined by the stick-breaking process (in contrast to the geometric rate associated with the Dirichlet process). This allows us to define the prior distribution of the number of distinct elements in a sample of n values drawn from the unknown distribution to be a heavier tailed distribution than under a Dirichlet process. The stick-breaking construction is defined by $V_j \sim \text{Be}(1 - a, M + aj)$ where $0 \leq a < 1$ and $M > -a$ and the Dirichlet process is the special case when $a = 0$. Again, case (i) of Theorem 1 applies and $w_{j,m} \sim \text{Be}(1 + M + a(m - 1), a)$ and $\epsilon_{j,m} = 0$ for $m \geq 1$.

Definition 4 *Let* $\tau_1, \tau_2, \tau_3, \ldots$ *follow a homogeneous Poisson process with intensity* λ. *We say that*

$\{G_t\}_{t=-\infty}^{\infty}$ *follows a PDAR*$(a, M, H; \lambda)$ *if*

$$G_t = \sum_{i=1}^{\infty} p_j(t)\delta_{\theta_j}$$

where $\theta_1, \theta_2, \theta_3, \cdots \overset{i.i.d.}{\sim} H$ *and*

$$p_j(t) = \begin{cases} 0 & \tau_j > t \\ V_j(t)\prod_{\{k|\tau_j<\tau_k<t\}}(1 - V_k(t)) & \tau_j < t \end{cases}$$

with

$$V_j(t) = \epsilon_j \prod_{i=2}^{m_j^{(t)}+1} w_{j,i}$$

where $\epsilon_j \sim$ *Be*$(1 - a, a + M)$ *and* $w_{j,m} \sim$ *Be*$(1 + M + a(m - 1), a)$ *for* $m > 1$.

This will always define a process with a Poisson-Dirichlet marginal. Figure (1) shows some realisations of the V_t process. For comparison, in the Dirichlet process case, $V_j(t)$ would be constant.

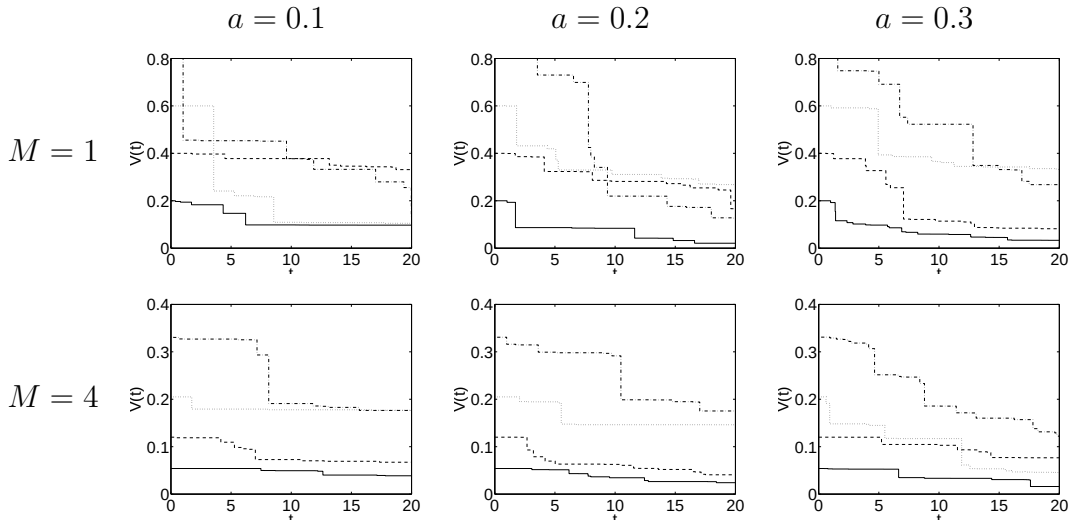

Figure 1: Four realisations of $V_j(t)$ for different values of a and M in the Poisson-Dirichlet process with $\lambda = 1$.

The effect of the Poisson-Dirichlet extension is to discount the value of $V_j(t)$ over time. As the weights of atoms are built up using factors $(1 - V_j(t))$ for the atoms that were introduced earlier, this leads to larger probabilities for atoms that were introduced in the past than under the Dirichlet process-generating scheme. As a increases the process for the $V_j(t)$ is increasingly discounted and

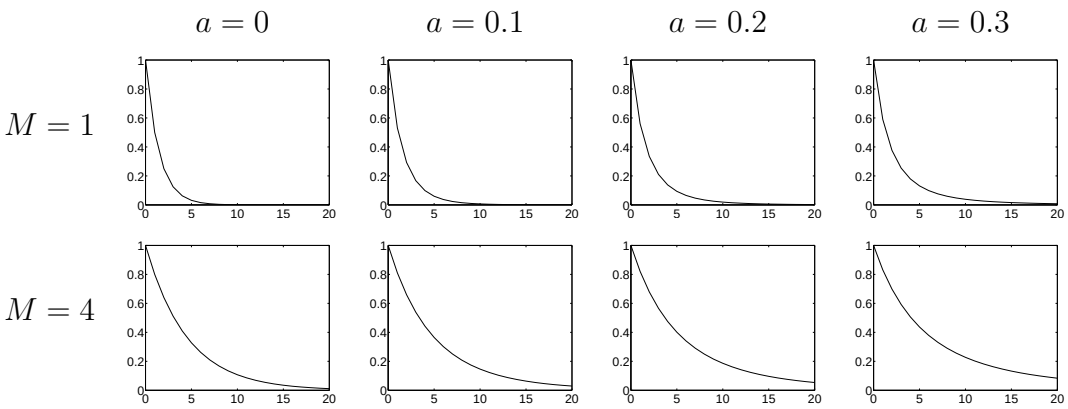

Figure 2: The autocorrelation function for various choice of a and M for a fixed value of λ

we generate larger autocorrelations for past shocks. This effect is illustrated by the shape of the autocorrelation function shown in Figure 2, which gives the autocorrelation function for various values of M and a and for a fixed value of λ. Larger values of a leads to increasingly slow decay of the autocorrelation function for fixed M. Larger values of M for fixed a lead to non-neglible autocorrelation at longer lags (as with the Dirichlet Process AR which corresponds to $a = 0$).

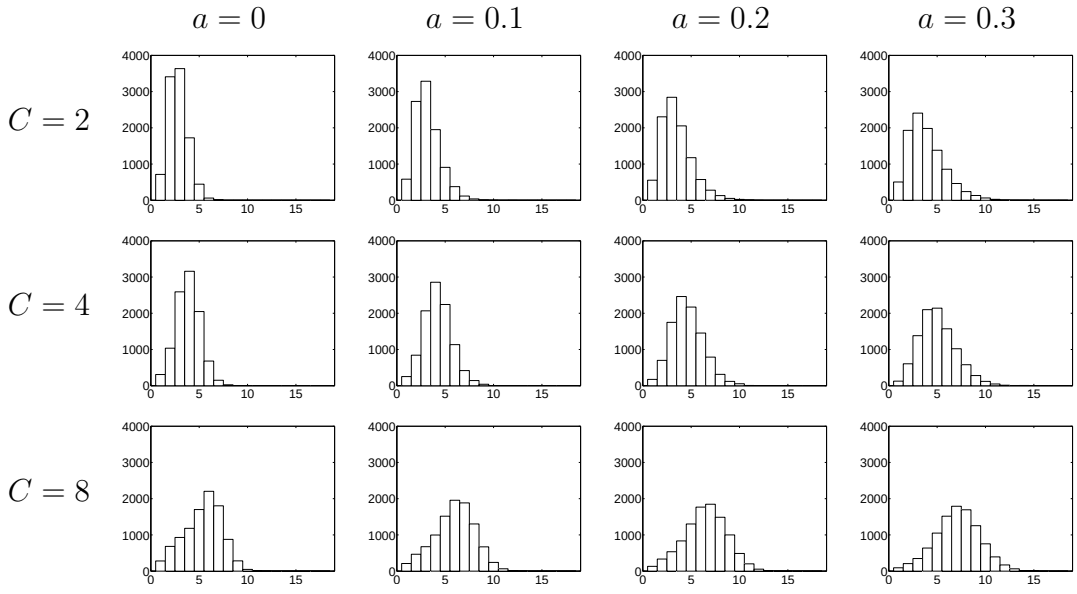

Figure 3: Number of clusters in 20 draws at time 2 conditional on C clusters in 20 draws at time 1: $M = 1$ and $\lambda = 1$

The Poisson-Dirichlet process gives us an extra parameter compared to the Dirichlet process which is related to the dispersion of the distribution of the number of clusters in a sample of size k drawn from the process. Larger values of a are related to a more dispersed distribution for given M. In our time series model, the parameter a controls the number of clusters in a sample of size k at time t given that there were C clusters at time $t-1$ in a sample of size k. Figure 3 shows the distribution for $k = 20$ and various values of M and C. For a fixed value of C, the mode of the distributions seems unchanged by the value of a but the right-hand tail has more mass on larger numbers of clusters. This provides the PDAR process with the extra flexiblity (over the Dirichlet process) to model the process when the number of clusters underlying the data is rapidly changing.

4 Computational methods

We will discuss Markov chain Monte Carlo methods for full sample inference. The MCMC sampler is defined when $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are both nondecreasing sequences. Samplers for Dirichlet process marginals can be defined using Pólya urn schemes developed in Section 3.1. Samplers for other marginal processes can be implemented by an extension of the retrospective sampler for Dirichlet processes (Papaspiliopoulos and Roberts, 2008). Appendix B groups most of the details; here we specifically focus on introducing latent variables that make Gibbs updating simpler by exploiting the stick-breaking form of $V_j(t)$ in this case. We assume that the arrival times of the atoms are $\tau_1 < \tau_2 < \tau_3 < \ldots$ and that associated with τ_j we have sequences $\epsilon_j = (\epsilon_{j,1}, \epsilon_{j,2}, \ldots)$ and $w_j = (w_{j,2}, w_{j,3}, \ldots)$. We will follow standard methods for stick-breaking mixtures by introducing latent variables $s_1, s_2, \ldots, s_n$ that allocate each observations to one of the atoms. We define r_i by $\tau_{r_i} = \max\{\tau_j | \tau_j < t_i\}$. Then, if $t_i > \tau_j$,

$$\begin{aligned} p(s_i = j) =& p_j(t_i) = V_j(t_i) \prod_{k=j+1}^{r_i} (1 - V_k(t_i)) \\ =& \left[\sum_{l=1}^{m_j(t_i)+1} \epsilon_{j,l}(1 - w_{j,l}) \prod_{h=l+1}^{m_j(t_i)+1} w_{j,h} \right] \prod_{k=j+1}^{r_i} \left[\sum_{l=1}^{m_k(t_i)+1} (1 - \epsilon_{k,l})(1 - w_{k,l}) \prod_{h=l+1}^{m_k(t_i)+1} w_{k,h} \right]. \end{aligned}$$

This form is not particular helpful for simulation of the posterior distribution but we define a more convenient form by introducing latent variables ϕ as follows

$$p(\phi_{ijk} = l) = (1 - w_{k,l}) \prod_{h=l+1}^{m_k(t_i)+1} w_{k,h}, \quad 1 \leq l \leq m_j(t_i) + 1,\ 1 \leq i \leq n,\ 1 \leq j \leq k \leq r_i$$

and

$$p(s_i = j|\phi) = \epsilon_{j,\phi_{ijj}} \prod_{k=j+1}^{r_i} (1 - \epsilon_{k,\phi_{ijk}}).$$

In effect, each stick-break is a mixture distribution and the indicators ϕ choose components of that mixture distribution. Then

$$p(s, \phi) = \prod_{i=1}^{n} \left[\epsilon_{s_i,\phi_{is_is_i}} \prod_{k=s_i+1}^{r_i} (1 - \epsilon_{k,\phi_{is_ik}}) \right] \prod_{i=1}^{n} \prod_{j=s_i}^{r_i} \prod_{k=j}^{r_i} \left[(1 - w_{k,\phi_{ijk}}) \prod_{h=\phi_{ijk}+1}^{m_j(t_i)+1} w_{k,h} \right].$$

which is a form that will be useful for simulation of the full conditional distributions of w and ϵ which will be beta distributed.

Further details on the MCMC algorithms proposed here are contained in Appendix B.

5 Empirical Results

5.1 Comparison of MCMC algorithms

We compare the marginal (see Section 3.1) and the general conditional algorithms with Dirichlet process marginals by analysing two simulated data sets and looking at the behaviour of the chain for the two parameters λ and M. The integrated autocorrelation time is used as a measure of the mixing of the two chains since an effective sample size can be estimated by sample size divided by integrated autocorrelation time (Liu 2001). We introduce three simple, simulated datasets to compare performance over a range of possible data.

In all cases, we make a single observation at each time point for $t = 1, 2, \ldots, 100$. The data sets are simulated from the following models. The first model has a single change point at time 50

$$p(y_i) = \begin{cases} \mathrm{N}(-20, 1) & \text{if } i < 50 \\ \mathrm{N}(20, 1) & \text{if } i \geq 50. \end{cases}$$

	M		λ	
Dataset	Cond.	Marg.	Cond.	Marg.
1	3.9	3.5	12.3	6.4
2	3.0	3.4	41.4	5.1
3	15.0	6.1	36.3	5.9

Table 1: The integrated autocorrelation times for M and λ using the two sampling schemes

The second model has a linear trend over time

$$p(y_i) = \mathrm{N}\left(\frac{40(i-1)}{99} - 20, 1\right)$$

The third model has a linear trend before time 40 and then follows a mixture of three regressions after time 40

$$p(y_i) = \begin{cases} \mathrm{N}\left(\frac{40(i-1)}{99} - 20, 1\right) & \text{if } i < 40 \\ \frac{3}{10}\mathrm{N}\left(\frac{40(i-1)}{99} - 20, 1\right) + \frac{2}{5}\mathrm{N}\left(-4, 1\right) + \frac{3}{10}\mathrm{N}\left(12 - \frac{40(i-1)}{99}, 1\right) & \text{if } i \geq 40. \end{cases}.$$

These data sets are fitted by a mixture of normals model

$$y_t \sim \mathrm{N}(\mu_t, 1)$$

$$\mu_t \sim G_t$$

$$G_t \sim DPAR(M, H; \lambda).$$

where $H(\mu) = \mathrm{N}(\mu|0, 100)$. Table 1 shows the results for the three data sets, using Exponential priors with unitary mean for M and λ. The mixing of λ is much better using the marginal sampler for each dataset (particularly data sets 2 and 3). The mixing of M is similar for the first two datasets but better for dataset 3. Thus, we use the marginal algorithm for the DPAR model in the following examples.

5.2 Real data examples

We consider three types of example. The first involves a long, single time series taken from the analysis of financial time series modelled by a stochastic volatility model with time-varying distri-

bution and the other two involve a short panel of many time series. In the latter cases we model the data through nonparametric mixtures of normal distributions. Throughout, the parameters M and λ were given exponential prior distributions with mean one.

5.2.1 Financial time series

Financial time series often show a number of stylized features such as heavy tails and volatility clustering (see *e.g.* Tsay 2005). Building models with stochastic volatility is a popular method for capturing these features. A stochastic volatility model assumes that the conditional variance follows a stochastic process. If $y_1, y_2, \ldots, y_T$ are the observations then a typical discrete time specification assumes that

$$y_t = \sqrt{h_t}\epsilon_t$$

where ϵ_t are i.i.d. from some returns distribution and the conditional variance h_t is modelled by

$$\log h_t \sim \mathrm{N}(\alpha + \delta \log h_{t-1}, \sigma_v^2).$$

The distribution of ϵ_t is usually taken to be normal. However, financial time series often contain large values which may not be fully explained by the time-varying variance. This has motivated the use of other choices. Jacquier *et al* (2004) considered using t-distributions. Jensen and Maheu (2007) consider a full Bayesian nonparametric model and show that the return distribution for an asset index may not be well-represented by either normal or t distributions. They model the returns distributions with a Dirichlet process mixture of normals. We extend the model by allowing the returns distribution to change over time. The hierarchical model can be written as

$$\frac{y_t}{\sqrt{h_t}} \sim \mathrm{N}(\mu_t, \sigma_t^2)$$

$$\log h_t \sim \mathrm{N}(\delta \log h_{t-1}, \sigma_v^2)$$

$$(\mu_t, \sigma_t^2) \sim G_t$$

$$G_t \sim DPAR(M, H; \lambda).$$

where $H\left(\mu, \sigma^{-2}\right) = \mathrm{N}(\mu|0, 0.01\sigma^2)\mathrm{Ga}\left(\sigma^{-2}\middle|\, 1, 0.1\right)$ and $\mathrm{Ga}(x|\alpha, \beta)$ represents a Gamma distribution with shape α and mean α/β. The parameters δ, σ_v^2 are given relatively flat priors for the

relevant ranges: $\sigma_v^{-2} \sim \text{Ga}(1, 0.005/2)$ and $\delta \sim \text{N}(0, 10)$ truncated to [0,1] as described in Jacquier *et al* (2004). Jensen and Maheu (2007) describe computational methods for the Dirichlet process-based model which can be extended using the method in Section 3.1. The method is applied to the daily returns of the Standard & Poors index from January 1, 1980 to December 30, 1987, as shown in Figure 4.

	Posterior median	95% HPD
M	0.37	(0.12,1.01)
ρ	0.997	(0.994, 0.999)

Table 2: Posterior inference for some parameters of the model

Posterior inference about the parameters of the model are shown in Table 2. The posterior median of M is around 0.37 indicating the nonparametric distribution has two or three normals which have non-neglible mass and the autocorrelation at 1 lag, ρ, is large showing that the distributions do not change rapidly over time. The posterior inference for the distributions is shown in Figure 4. As suggested by the estimates of M and ρ we have results that would be roughly consistent with a changepoint analysis where several regions with very similar returns distributions have been identified. Figure 4(b) shows representative distributions for the main periods and illustrates the range of shapes in the returns distribution. The main difference between the distributions is their spread and the variance of the fitted distributions is shown in Figure 4(c). The results are extremely smooth and can be thought of as representing an estimate of underlying, long-run volatility (since daily changes in volatility are captured through the volatility equation). A parametric analysis assuming that the returns distribution is normal leads to an estimate of the long-run variance to be 1.66 which is roughly an average of our nonparametric estimates.

5.2.2 Income data

The data contain the real (log) per capita GDP of 110 EU regions from 1977 to 1996 and has been previously analysed by Grazia Pittau and Zelli (2006). We ignore the longitudinal nature of the data and assume that the problem can be treated as density estimation over time with independent

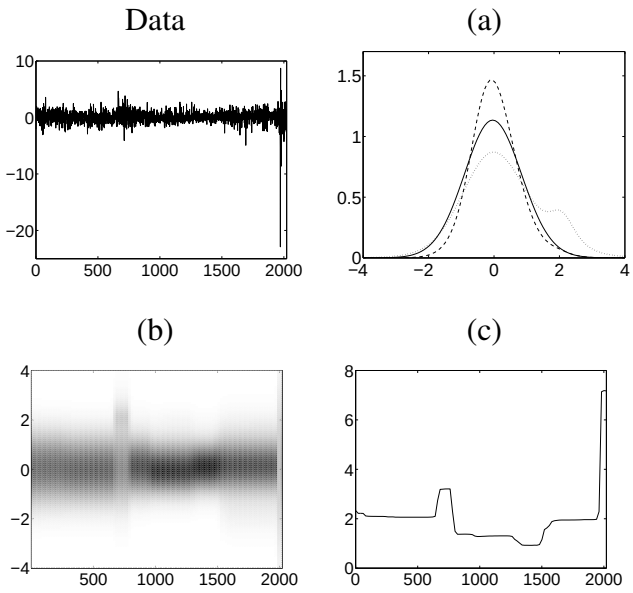

Figure 4: Inference for the Standard & Poors data set: (a) selected predictive density functions for observables at times 481, 681 and 1181; (b) heatmap of the predictive density functions of the returns at each time point (darker colours represent higher density values); (c) variance of the fitted returns distribution over time

measurements and model

$$y_{it} \sim \mathrm{N}(\mu_{it}, \sigma_{it}^2)$$

$$(\mu_{it}, \sigma_{it}^2) \sim G_t$$

$$G_t \sim DPAR(M, H; \lambda) \tag{7}$$

where $H\left(\mu, \sigma^{-2}\right) = \mathrm{N}(\mu|\mu_0, 0.01\sigma^2)\mathrm{Ga}\left(\sigma^{-2}\middle|\, 1, 0.1\right)$. The hyperparameter μ_0 represents an overall mean value for the data and the sample mean is adopted as a suitable value. Results are presented in figure 5. Panel (a) shows a heatmap of the estimated distribution plotted at each year. The most striking feature of the plots is that the distribution changes from year 1988 to 1989 with larger values observed from 1989 onwards. It is clear the model very much behaves like a change-point model with one change-point. Panel (b) shows the estimated densities for each year. The change in the main model of the data is obvious but there is also a change in the modality of the data and the shape of the tails. To check whether this change in distribution is supported by the data, we fitted independent Dirichlet process mixture models to each year. The results are presented in panel (c)

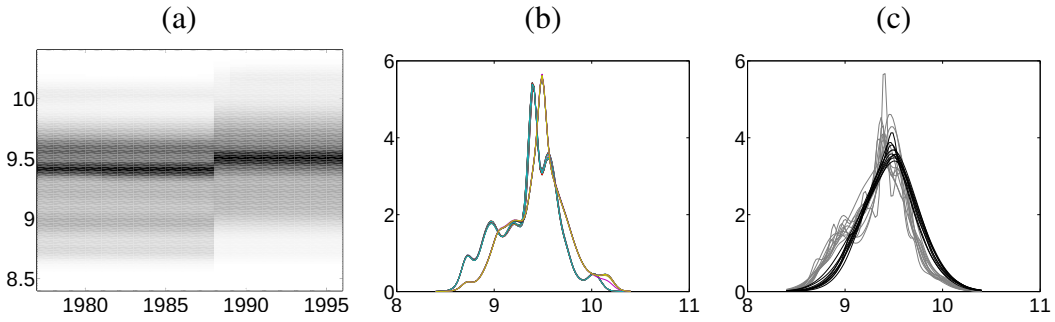

Figure 5: Income data: (a) heatmap of the estimated distribution for each year using a DPAR mixture model; (b) density estimates for each year using DPAR; (c) density estimates for each year using independent Dirichlet process mixture models (pre-1989 shown in light grey and other years in black).

and support the two main distributions inferred from the data. In fact, the yearly distributions are very similar for the second period (post-1988). It is interesting to note that the density estimates are much smoother for the independent compared to the DPAR model, which is due to the smaller amount of information available for each estimate.

5.2.3 NUTS data

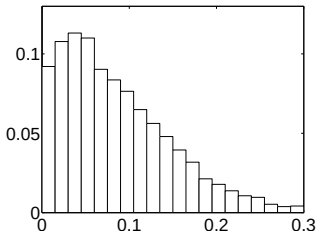

Figure 6: Posterior distribution of a for the NUTS data

The data consists of annual per capita GDP growth rates for 258 NUTS2 European regions covering the period from 1995 to 2004 (NUTS2 regions provide a roughly uniform breakdown of the EU into territorial units). The data are modelled using a mixture of normal distributions as in model (7) with the exception that now we use the model with Poisson-Dirichlet marginals: $G_t \sim \text{PDAR}(a, M, H; \lambda)$. We consider the cases where a is fixed and where a is given a uniform prior distribution on $[0, 1)$.

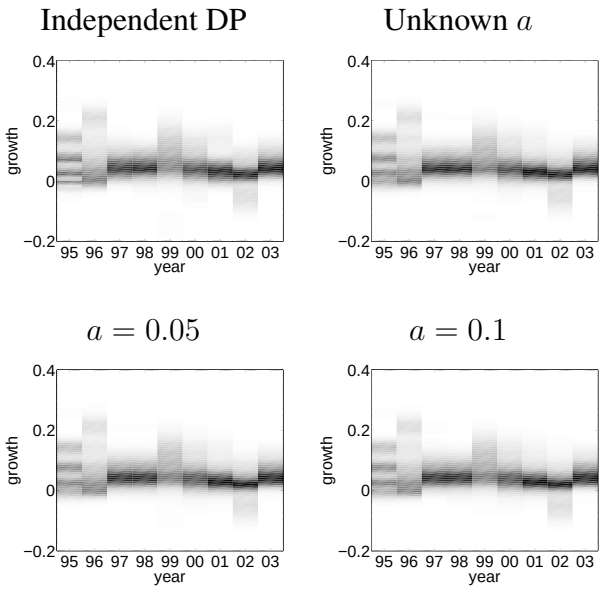

Figure 7: Heatmaps of the fitted yearly distribution of growth for the NUTS data

Figure 6 shows the posterior distribution of a, which places its mass on smaller values of a (under 0.3) with mode around 0.05. The yearly predictive distribution of growth is shown in figure 7 for the PDAR with $a = 0.05$ and $a = 0.1$ and a unknown. This figure also presents results for a model where the distribution of each year's growth is estimated independently with a Dirichlet process mixture. The results are remarkably consistent across all models. This is perhaps not surprising since we are looking at posterior means with a substantial amount of data in the sample and large differences between each year's distribution. When the data is thinned at random to a

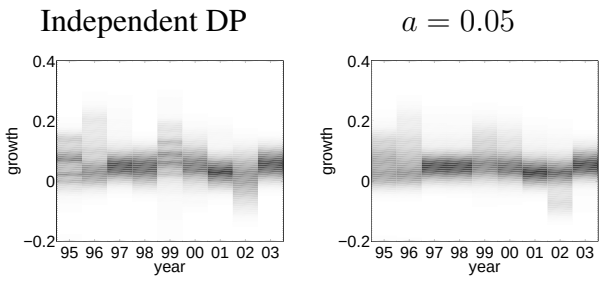

Figure 8: Heatmaps of the fitted yearly distribution of growth for the thinned NUTS data

sample of 60 regions over 9 years, Figure 8 contrasts the results for independent DP and the PDAR model with $a = 0.05$. The results for the PDAR model then show more smoothing, particularly

when distributions in consecutive years are similar, as we would expect from a "change-point" type analysis. However, even with the full data set there are differences between the posterior distributions of the parameters of the model. Figure 9 shows the posterior distribution of λ. This is

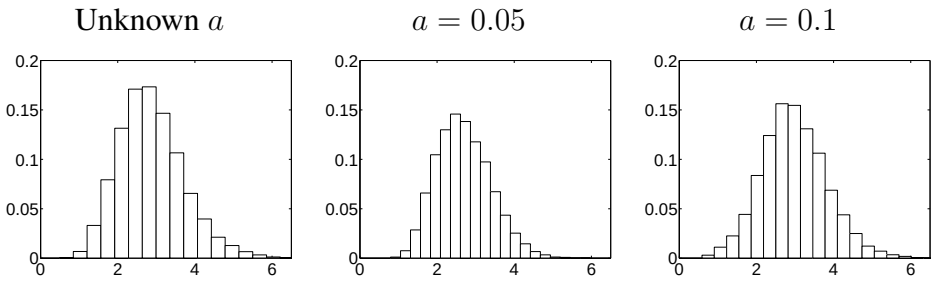

Figure 9: Posterior distribution of λ for the NUTS data

the mean number of new clusters introduced each year. The distribution is concentrated between 2 and 4. Once again, this indicates the large differences between the distributions for each year. The mean of λ is 2.69 when $a = 0.05$ and 2.98 when $a = 0.1$. When a is unknown the mean, 2.87, falls between these two values. As we increase a in the Poisson-Dirichlet model then we are more likely to introduce smaller components which allows the introduction of larger numbers of components at each year. This idea is supported by the posterior distribution of the number of clusters: the median

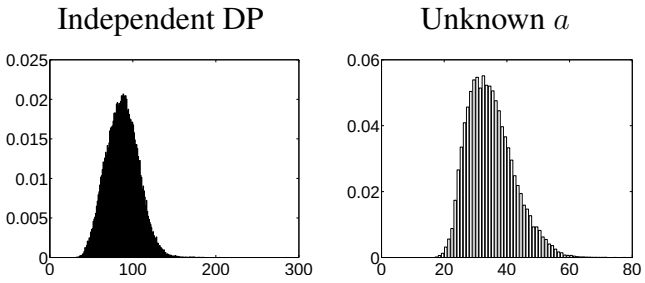

Figure 10: Posterior distribution of the number of clusters for the NUTS data

number of clusters is 32 if $a = 0.05$ and 36 if $a = 0.1$. When a is unknown the median number of clusters is 34. Figure 10 displays its posterior distribution and also presents the posterior distribution of the number of clusters when we model the data in each year with independent Dirichlet process mixture models. In the latter case we obtain a substantially larger number of clusters (the median is 88, *i.e.* roughly two-and-a-half times the number under the time-dependent model). This suggests

that despite the lack of similarity between the distribution of each year some clusters can usefully be carried over from one year to the next.

6 Discussion

This paper introduces, develops and implements inference with a new class of time-dependent measure-valued processes with stick-breaking marginals, which can be used as a prior distribution in nonparametric time-series modelling. The Dirichlet process and Poisson-Dirichlet process marginals arise as natural special cases. We derive a Pólya urn scheme for the Dirichlet process case which allows us to develop a new algorithm using a marginalised method. This method typically leads to better mixing of the parameter (particularly the intensity parameter of the Poisson process). We also develop a conditional simulation method using retrospective sampling methods when the parameters of the stick-breaking process are nondecreasing.

Moving from Dirichlet process to Poisson-Dirichlet process marginals allows us to more closely control the conditional distribution of the number of clusters in a sample of size n at time t given the number at time $t-1$. The processes provide smoothed estimates of the distributions of interest. The models can behave like a change point model which allows the discovery of periods where distributions are relatively unchanged.

References

Antoniak, C. E. (1974): "Mixtures of Dirichlet processes with applications to non-parametric problems," *Journal of the American Statistical Association*, 2, 1152-1174.

Carlin, B. P., A. E. Gelfand and A. F. M. Smith (1992): "Hierarchical Bayesian analysis of change-point problems," *Applied Statistics*, 41, 389-405.

Caron, F., M. Davy and A. Doucet (2007): "Generalized Pólya Urn for Time-varying Dirichlet Process Mixtures", *23rd Conference on Uncertainty in Artificial Intelligence (UAI 2007).*

De Iorio, M., P. Müller, G. L. Rosner and S. N. MacEachern (2004): "An ANOVA model for dependent random measures, " *Journal of the American Statistical Association*, 99, 205-215.

Doucet, A., N. de Freitas and N. J. Gordon (2001): "Sequential Monte Carlo Methods in Practice," Springer-Verlag: New York.

Dunson, D. B. (2006): "Bayesian dynamic modeling of latent trait distributions," *Biostatistics*, 7, 551-568.

Dunson, D. B., N. Pillai and J. H. Park (2007): "Bayesian density regression, " *Journal of the Royal Statistical Society B*, 69, 163-183.

Grazia Pittau, M. and R. Zelli (2006): "Empirical Evidence of Income Dynamics Across EU Regions," *Journal of Applied Econometrics*, 21, 605-628.

Griffin, J. E. (2007): "The Ornstein-Uhlenbeck Dirichlet Process and other time-varying processes for Bayesian nonparametric inference," Working Paper 07-03, CRiSM, University of Warwick.

Griffin, J. E. and M. F. J. Steel (2006): "Order-based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179-194.

Ishwaran, H. and L. F. James (2001): "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161-173.

Ishwaran, H. and L. F. James (2003). "Some further developments for stick-breaking priors: finite and infinite clustering and classification," *Sankhya, A*, 65, 577-592.

Ishwaran, H. and M. Zarepour (2000): "Markov chain Monte Carlo in approximate Dirichlet and two-parameter process hierarchical models," *Biometrika*, 87, 371-390.

Jacquier, E., N. G. Polson and P. E. Rossi (2004): "Bayesian analysis of stochastic volatility models with fat tails and correlated errors," *Journal of Econometrics*, 122, 185-212.

James, L.F. (2008): "Large sample asymptotics for the two-parameter PoissonDirichlet process," in *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, B. Clarke and S. Ghosal, eds., IMS: Beachwood, 187-199.

James, L. F., A. Lijoi and I. Prünster (2005): "Bayesian inference via classes of normalized random measures," Technical Report.

Jensen, M. J. and J. M. Maheu (2007): "Bayesian semiparametric stochastic volatility modeling," Technical Report, University of Toronto.

Liu, J. S. (2001): *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag: New York.

Lo, A. Y. (1984): "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12, 351-357.

Müller, P., F. Quintana, and G. Rosner (2004): "A method for combining inference across related nonparametric Bayesian models," *Journal of the Royal Statistical Society B*, 66, 735-749.

Nieto-Barajas, L., Müller, P., Ji, Y., Lu, Y. and Mills, G. (2008): "Time Series Dependent Dirichlet Process," *Preprint*

Papaspiliopoulos, O. and G. Roberts (2008): "Retrospective MCMC for Dirichlet process hierarchical models," *Biometrika*, 95, 169-186.

Pitman, J. (1996): "Some Developments of the Blackwell-MacQueen Urn Scheme," in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, eds: T. S. Ferguson, L. S. Shapley and J. B. MacQueen, IMS Lecture Notes.

Pitman, J. and M. Yor (1997): "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, 25, 855-900.

Rodriguez, A. and E. ter Horst (2008): "Bayesian Dynamic Density Estimation," *Bayesian Analysis*, 3, 339-366.

Teh, Y. W. (2006): 'A Hierarchical Bayesian Language Model based on Pitman-Yor Processes," *Proceedings of the Annual Meeting of the Association for Computational Linguistics 44*.

Tsay, R. (2005): *Analysis of Financial Time Series (2nd Edition)*, John Wiley & Sons: New York.

Zhu, X., Z. Ghahramani, and J. Lafferty (2005): "Time-Sensitive Dirichlet Process Mixture Models," Technical Report CMU-CALD-05-104, Carnegie Mellon University.

A Appendix

A.1 Proof of Theorem 1

These proofs use the following properties:

1. Suppose that $q \sim \text{Ga}(a)$ and $r \sim \text{Ga}(b)$ are independent then $v = \frac{q}{q+r} \sim \text{Be}(a, b)$ which is independent of $u = q + r \sim \text{Ga}(a + b)$.
2. If $q \sim \text{Ga}(a + b)$ and independent of $v \sim \text{Be}(a, b)$ then $vq \sim \text{Ga}(a)$ and independent of $(1 - v)q \sim \text{Ga}(b)$.

Property 1 implies that any beta random variable, v, can be expressed as $\frac{q}{q+r}$ where q and r are independent. We can write $V_{j,t} = \frac{q_{j,t}}{q_{j,t}+r_{j,t}}$, where $q_{j,t} \sim \text{Ga}(a_t)$ and $r_{j,t} \sim \text{Ga}(b_t)$ for all cases.

Case (i) Let $q_{j,t+1} = q_{j,t} + x_{j,t+1}$ and $r_{j,t+1} = r_{j,t} + z_{j,t+1}$ where $x_{j,t+1} \sim \text{Ga}(a_{t+1} - a_t)$ and independent of $z_{j,t+1} \sim \text{Ga}(b_{t+1} - b_t)$. Then $q_{j,t+1} \sim \text{Ga}(a_{t+1})$ and $r_{j,t+1} \sim \text{Ga}(b_{t+1})$ and $q_{j,t+1}$ and $r_{j,t+1}$ are independent. Then $V_{j,t+1} = \frac{q_{j,t+1}}{q_{j,t+1}+r_{j,t+1}}$ is beta distributed with the correct parameters. Simple algebra shows that we can write

$$V_{j,t+1} = \frac{q_{j,t+1}}{q_{j,t+1} + r_{j,t+1}} = w_{j,t+1}V_{j,t} + (1 - w_{j,t+1})\epsilon_{j,t+1}$$

where $w_{j,t+1} = \frac{q_{j,t}+r_{j,t}}{q_{j,t}+r_{j,t}+x_{j,t}+z_{j,t}}$ and $\epsilon_{j,t+1} = \frac{x_{j,t}}{x_{j,t}+z_{j,t}}$. From the property of beta-gamma distributions $x_{j,t} + z_{j,t}$ is independent of $\epsilon_{j,t+1}$ and it follows that $w_{j,t+1}$ is independent of $\epsilon_{j,t+1}$. Finally, $w_{j,t+1} \sim \text{Be}(a_{t+1} + b_{t+1}, a_{t+1} + b_{t+1} - a_t - b_t)$ and $\epsilon_{j,t+1} \sim \text{Be}(a_{t+1} - a_t, b_{t+1} - b_t)$

Case (ii) Then we can write $q_{j,t+1} = u_{j,t+1}q_{j,t}$ and $r_{j,t+1} = y_{j,t+1}r_{j,t}$ where $u_{j,t+1} \sim \text{Be}(a_{t+1}, a_t - a_{t+1})$ and $y_{j,t+1} \sim \text{Be}(b_{t+1}, b_t - b_{t+1})$. Application property 2 shows that $q_{j,t+1}\text{Ga}(a_{t+1})$ and $r_{j,t+1} \sim \text{Ga}(b_{t+1})$ are independent and so $V_{j,t+1} = \frac{q_{j,t+1}}{q_{j,t+1}+r_{j,t+1}}$ is beta distributed with the correct parameters. This can be re-expressed as follows

$$V_{j,t+1} = \frac{q_{j,t+1}}{q_{j,t+1} + r_{j,t+1}} = \frac{w_{j,t+1}V_{j,t}}{w_{j,t+1}V_{j,t} + (1 - w_{j,t+1})(1 - V_{j,t})}$$

where $w_{j,t+1} = \frac{u_{j,t+1}}{u_{j,t+1}+y_{j,t+1}}$.

Case (iii) Then we can write $q_{j,t+1} = y_{j,t+1}q_{j,t}$ and $r_{j,t+1} = r_{j,t} + z_{j,t+1}$ where $u_{j,t+1} \sim \text{Be}(a_{t+1}, a_t - a_{t+1})$ and $z_{j,t+1} \sim \text{Ga}(b_{t+1} - b_t)$.

$$V_{j,t+1} = \frac{q_{j,t+1}}{q_{j,t+1} + r_{j,t+1}} = \frac{u_{j,t+1}q_{j,t}}{u_{j,t+1}q_{j,t} + r_{j,t} + z_{j,t+1}}.$$

Case (iv) Then we can write $q_{j,t+1} = q_{j,t} + x_{j,t+1}$ and $r_{j,t+1} = y_{j,t+1}r_{j,t}$ where $x_{j,t+1} \sim \text{Ga}(a_{t+1} - a_t)$ and $y_{j,t+1} \sim \text{Be}(b_{t+1}, b_t - b_{t+1})$.

$$V_{j,t+1} = \frac{q_{j,t+1}}{q_{j,t+1} + r_{j,t+1}} = \frac{q_{j,t} + x_{j,t+1}}{x_{j,t+1} + q_{j,t} + y_{j,t+1}r_{j,t}}.$$

A.2 Proof of Theorem 2

Let $k_i = \#\{i|\phi_i < \tau_i < \phi_{i+1}\}$ for $1 \leq i \leq l_{n,m} - 1$.

$$p(k_1, k_2, \ldots, k_{l_{n,m}-1}|s_1, s_2, \ldots, s_n) \propto \prod_{i=1}^{l_{n,m}-1} \left(\frac{M}{M + A_n(\phi_{i+1}}\right)^{k_i} (\lambda(\phi_{i+1} - \phi_i))^{k_i}$$

which shows that the number of points on (ϕ_i, ϕ_{i+1}) is Poisson distributed with mean $\left(\frac{M}{M+A_n(\phi_{i+1}}\right)\lambda(\phi_{i+1} - \phi_i)$. The position of the points is unaffected by the likelihood and so the posterior is a Poisson process. There is no likelihood contribution for the intervals $(-\infty, \phi_1)$ and $(\phi_{l_{n,m}}, \infty)$. Since the Poisson process has independent increment then the posterior distribution on these intervals is also a Poisson process with intensity λ.

A.3 Proof of Theorem 3

In order to calculate the predictive distribution we need to calculate the probability of generating the sample $s_1, s_2, \ldots, s_n$ which given by

$$p(s_1, s_2, \ldots, s_n) = \mathrm{E}[p(s_1, \ldots, s_n|V_1, V_2, V_3, \ldots, \tau_1, \tau_2, \tau_3, \ldots)].$$

This expectation can be derived by first noting that

$$p(s_1, \ldots, s_n|V_1, V_2, V_3, \ldots, \tau_1, \tau_2, \tau_3, \ldots) = \prod_{i \in \mathcal{R}} V_i^{\eta_i}(1 - V_i)^{A_n(\tau_i)}$$

where $\mathcal{R} = \{i| \min\{\tau_{s_i}, 1 \leq i \leq n\} \leq \tau_i \leq \max\{t_i, 1 \leq i \leq n\}\}$. Marginalising over V gives

$$p(s_1, \ldots, s_n|\tau_1, \tau_2, \tau_3, \ldots) = \prod_{i \in \mathcal{R}} \frac{M\eta_i!\Gamma(M + A_n(\tau_i))}{\Gamma(M + 1 + \eta_i + A_n(\tau_i))}.$$

Noticing that, if $\eta_i = 0$ then

$$\frac{M\eta_i!\Gamma(M + A_n(\tau_i))}{\Gamma(M + 1 + \eta_i + A_n(\tau_i))} = \frac{M}{M + A_n(\tau_i)},$$

it follows that

$$p(s_1, \ldots, s_n|\mathcal{S}_{n,m}) = \prod_{i=1}^{k_n} \frac{M\eta_i!\Gamma(M + A_n(\tau_i^\star))}{\Gamma(M + 1 + \eta_i + A_n(\tau_i^\star))} \prod_{i=2}^{l_{n,m}} \mathrm{E}\left[\left(\frac{M}{M + A_n(\phi_i)}\right)^{\#\{j|\phi_{i-1}<\tau_j<\phi_i\}}\right].$$

From Theorem 2, $\#\{j|\phi_{i-1} < \tau_j < \phi_i\}$ is Poisson distributed with mean $\lambda\left(\frac{M}{M+A_n(\phi_i)}\right)(\phi_i - \phi_{i-1})$ and so

$$\mathrm{E}\left[\left(\frac{M}{M + A_n(\phi_i)}\right)^{\#\{j|\phi_{i-1}<\tau_j<\phi_i\}}\right] = \exp\left\{-\lambda(\phi_i - \phi_{i-1})\frac{MA_n(\phi_i)}{(M + A_n(\phi_i))^2}\right\}$$

and

$$p(s_1,\ldots,s_n|\mathcal{S}_{n,m}) = \prod_{i=1}^{k_n} \frac{M\eta_i!\Gamma(M+A_n(\tau_i^\star))}{\Gamma(M+1+\eta_i+A_n(\tau_i^\star))} \prod_{i=2}^{l_{n,m}} \exp\left\{-\lambda(\phi_i-\phi_{i-1})\frac{MA_n(\phi_i)}{(M+A_n(\phi_i))^2}\right\}$$

it also follows that if $s_{n+1}=j$ where $j\leq k_n$ then, if $m\geq n+1$,

$$\begin{aligned} p(s_1,\ldots,s_n,s_{n+1}=j|\mathcal{S}_{n,m}) =& \frac{M(\eta_j+1)!\Gamma(M+A_n(\tau_i^\star))}{\Gamma(M+2+\eta_j+A_n(\tau_i^\star))} \prod_{i=1;i\neq j}^{k_n} \frac{M\eta_i!\Gamma(M+A_{n+1}(\tau_i^\star))}{\Gamma(M+1+\eta_i+A_{n+1}(\tau_i^\star))} \\ & \times \prod_{i=2}^{l_{m,n}} \exp\left\{-\lambda(\phi_i-\phi_{i-1})\left(\frac{M}{M+A_n(\phi_i)}\right)\frac{A_{n+1}(\phi_i)}{M+A_{n+1}(\phi_i)}\right\}. \end{aligned}$$

So that, if $j\leq k_n$,

$$p(s_{n+1}=j|\mathcal{S}_{n,m},s_1,\ldots,s_n) = \frac{p(s_1,\ldots,s_n,s_{n+1}=j|\mathcal{S}_{n,m})}{p(s_1,\ldots,s_n|\mathcal{S}_{n,m})};$$

after some algebra we get the form in the Theorem. Otherwise, $s_{n+1}=k_n+1$ and we need to calculate $p(s_1,\ldots,s_n,s_{n+1}=k_n+1,\tau^\star_{k_n+1}\in(\phi_{i-1},\phi_i)|\mathcal{S}_{n,m})$. It is clear that

$$\frac{M\eta_i!\Gamma(M+A_{n+1}(\phi_i))}{\Gamma(M+1+\eta_i+A_{n+1}(\phi_i))} = \begin{cases} \frac{M}{M+A_n(\phi_i)} & \text{if } \phi_{i-1}<\tau_i<\tau^\star_{k_n+1} \\ \frac{M}{(M+A_n(\phi_i))(M+1+A+n(\phi_i))} & \text{if } \tau_i=\tau^\star_{k_n+1} \\ \frac{M}{M+A_n(\phi_i)+1} & \text{if } \tau^\star_{k_n+1}<\tau_i<\phi_i \end{cases}.$$

Let $\tau_{k_n-1}=(1-w)\phi_i+w\phi_{i-1}$,

$$\begin{aligned} & \mathrm{E}\left[\left(\frac{M}{M+A_{n+1}(\phi_i)}\right)^{\#\{j|\phi_{i-1}<\tau_j<\phi_i\}}\middle|\tau^\star_{k_n+1},\#\{j|\phi_{i-1}<\tau_j<\phi_i\}=k\right] \\ =& \frac{M}{(M+A_n(\phi_i))(M+1+A_n(\phi_i))}k\left[\frac{M}{M+A_n(\phi_i)+1}w+\frac{M}{M+A_n(\phi_i)}(1-w)\right]^{k-1} \end{aligned}$$

and so

$$\begin{aligned} & \mathrm{E}\left[\left(\frac{M}{M+A_{n+1}(\phi_i)}^{\#\{j|\phi_i<\tau_j<\phi_{i+1}\}}\right)\middle|\tau^\star_{k_n+1}\right] \\ =& \frac{M}{(M+A_n(\phi_i))(M+1+A_n(\phi_i))} \\ \times& \frac{\lambda M(\phi_i-\phi_{i-1})}{M+A_n(\phi_i)}\left[\exp\left\{-\frac{\lambda M(\phi_i-\phi_{i-1})}{M+A_n(\phi_i)}\left[\frac{A_n(\phi_i)+1}{M+A_n(\phi_i)+1}w+\frac{A_n(\phi_i)}{M+A_n(\phi_i)}(1-w)\right]\right\}\right]. \end{aligned}$$

Finally $\tau^\star_{k_n+1}$ is uniformly distributed on (ϕ_{i-1},ϕ_i) which implies that

$$\begin{aligned} & \mathrm{E}\left[\left(\frac{M}{M+A_n(\phi_i)}^{\#\{j|\phi_{i-1}<\tau_j<\phi_i\}}\right)\right] \\ =& \exp\left\{-\frac{M\lambda(\phi_i-\phi_{i-1})A_n(\phi_i)}{(M+A_n(\phi_i))^2}\right\}\left[1-\exp\left\{-\frac{M^2\lambda(\phi_i-\phi_{i-1})}{(M+A_n(\phi_i))^2(M+A_n(\phi_i)+1)}\right\}\right]. \end{aligned}$$

It follows that we can write

$$p(s_1,\ldots,s_n,s_{n+1}\in(\phi_{j-1},\phi_j)|\mathcal{S}_{n,m}) = \prod_{i=1}^{k_n}\frac{M\eta_i!\Gamma(M+A_{n+1}(\tau_i^\star))}{\Gamma(M+1+\eta_i+A_{n+1}(\tau_i^\star))}\exp\left\{-\frac{M\lambda(\phi_j-\phi_{j-1})A_n(\phi_j)}{(M+A_n(\phi_j))^2}\right\}$$
$$\times\left[1-\exp\left\{-\frac{M^2\lambda(\phi_j-\phi_{j-1})}{(M+A_n(\phi_j))^2(M+A_n(\phi_j)+1)}\right\}\right]$$
$$\times\prod_{i=2;i\neq j}^{l_{m,n}}\exp\left\{-\lambda(\phi_i-\phi_{i-1})\left(\frac{M}{M+A_n(\phi_i)}\right)\frac{A_{n+1}(\phi_i)}{M+A_{n+1}(\phi_i)}\right\}$$

and

$$p(s_{n+1}=k_n+1,\tau^\star_{k_n+1}\in(\phi_{i-1},\phi_i)|\mathcal{S}_{n,m},s_1,\ldots,s_n)$$
$$=\frac{p(s_1,\ldots,s_n,s_{n+1}=k_n+1,\tau^\star_{k_n+1}\in(\phi_{i-1},\phi_i)|\mathcal{S}_{n,m})}{p(s_1,\ldots,s_n|\mathcal{S}_{n,m})}.$$

Finally, since τ_{k_n+1} is uniformly distributed on (ϕ_i,ϕ_{i-1}) if $\tau_{k_n+1}\in(\phi_i,\phi_{i-1})$ then

$$p(\tau^\star_{k_n+1}=\phi_i-x|\tau^\star_{k_n+1}\in(\phi_{i-1},\phi_i))$$
$$\propto \mathrm{E}\left[\left(\frac{M}{M+A_n(\phi_i)}\right)^{\#\{j|\phi_{i-1}<\tau_j<\phi_i\}}\right]$$
$$\propto \exp\left\{-\frac{\lambda M^2}{(M+A_n(\phi_i))^2(M+A_n(\phi)+1)}(\phi_i-\tau_{k_n+1})\right\}$$

which implies that $x=\phi_i-\tau_{k_n+1}$ follows the distribution given in the Theorem.

B Computational Details

We will write the times in reverse time-order $T>\tau_1>\tau_2>\tau_3>\cdots>\tau_k$ where $T=\max\{t_i\}$ and $k=\max\{s_j\}$. Let $k_{-i}=\max_{j\neq i}\{s_j\}$. We will use the notation from Definition 2, $m_j(t)=\#\{k|\tau_j<\tau_k<t\}$.

B.1 General sampler

Updating s

We update s_i using a retrospective sampler (see Papaspiliopoulos and Roberts 2008). Let $\Delta=\{\theta,w,\epsilon,\tau\}$. This method proposes a new value of (s_i,Δ), which will be referred to as (s_i',Δ'), that are either accepted or rejected in a Metropolis-Hastings sampler. The proposal is made in the following way:

Let $\theta_i'=\theta_i$, $\epsilon_i'=\epsilon_i$, $w_i'=w_i$ and $\tau_i'=\tau_i$ for $1\leq i\leq k_{-i}$, $\alpha=\max_{j\leq k_{-i}}\{k(y_i|\theta_j')\}$ and define

$$q_j=\frac{p(s_i=j)k(y_i|\theta_j')}{\alpha\left(1-\sum_{j=1}^{k_{-i}}p(s_i=j)\right)+\sum p(s_i=j)k(y_i|\theta_j')},\qquad 1\leq j\leq k_{-i}.$$

Simulate $u \sim \mathrm{U}(0,1)$. If $u < \sum_{j=1}^{k_{-i}} q_j$ then find the m for which $\sum_{j=1}^{m-1} q_j < u < \sum_{j=1}^{m} q_j$. Otherwise, we simulate in the following way. Let

$$q_j = \alpha p(s_i = j), \qquad j > k_{-i}$$

and sequentially simulate $\Delta'_{k_{-i}+1}, \Delta'_{k_{-i}+2}, \ldots, \Delta'_m$ until we meet the condition that $u < \sum_{j=1}^{m} q_j$. We can simulate Δ'_j given Δ'_{j-1} using the relation $\tau'_j = \tau'_{j-1} - \nu_j$ where $\nu_j \sim \mathrm{Ex}(\lambda)$ and simulating θ'_j, ϵ'_j and w'_j from their prior. The new state (s'_i, Δ') is accepted with probability

$$\zeta = \begin{cases} 1 & \text{if } m \leq k_{-i} \\ \min\left\{1, \frac{k(y_i|\theta'_m)}{\alpha}\right\} & \text{if } m > k_{-i} \end{cases}.$$

Updating ϕ

If $j = s_i$, then the full conditional distribution of ϕ_{ijk} is given by

$$p(\phi_{is_is_i} = l) \propto \epsilon_{s_i,l}(1 - w_{j,l}) \prod_{h=l+1}^{m_j(t_i)} w_{j,h}$$

$$p(\phi_{is_ik} = l) = (1 - \epsilon_{s_i,l})(1 - w_{j,l}) \prod_{h=l+1}^{m_j(t_i)} w_{j,h}, \quad k < s_i$$

Otherwise ϕ_{ijk} is sampled from its prior distribution.

Updating ϵ

The full conditional distribution of $\epsilon_{j,k}$ is

$$\mathrm{Be}\left(a_k - a_{k-1} + \sum_{\{i|s_i=j\}} \mathrm{I}(\phi_{ijj} = k), b_k - b_{k-1} + \sum_{\{i|s_i<j \text{ and } \tau_j<t_i\}} \sum_{p=s_i+1}^{r_i} \mathrm{I}(\phi_{ijp} = k)\right).$$

Updating w

The full conditional distribution of $w_{j,l}$ is $\mathrm{Be}\left(a^\star, b^\star\right)$, where

$$a^\star = a_{l-1} + b_{l-1} + \sum_{i=1}^{n} \sum_{j=s_i}^{r_i} \sum_{k=j}^{r_i} \mathrm{I}(\phi_{ijk} + 1 \leq l \leq m_j(t_i))$$

and

$$b^\star = a_l + b_l - a_{l-1} - b_{l-1} + \sum_{h=s_i}^{r_i} \sum_{i=1}^{n} \mathrm{I}(h \leq j \leq r_i)\mathrm{I}(\phi_{ihj} = l).$$

Updating τ

The point process τ can be updated using a Reversible Jump MCMC step. We have three possible move: 1) Add a point to the process, 2) delete a point from the process and 3) Move a point. The

first two moves are proposed with the same probability q_{CHANGE} (where $q_{CHANGE} < 0.5$) and the third move is proposed with probability $1 - 2q_{CHANGE}$. The Add move proposes the addition of a point to the process by uniformly sampling τ_{k+1} from $(\min\{\tau_i\}, \max\{t_i\})$, $\theta_{k+1} \sim H$ and simulating the necessary extra ϵ's, w's and ϕ's from their prior. To improve acceptance rates we also update some allocations s. A point, $j^\star$, is chosen uniformly at random from $\{1, \ldots, k\}$ and we propose new values s_i' if $s_i = j^\star$ according to the probabilities

$$p(s_i' = j^\star) = q_{i,1} = \frac{p'(s_i = j^\star)k\left(y_i|\theta_{j^\star}\right)}{p'(s_i = k+1)k\left(y_i|\theta_{k+1}\right) + p'(s_i = j^\star)k\left(y_i|\theta_{j^\star}\right)}$$

and

$$p(s_i' = k+1) = q_{i,2} = \frac{p'(s_i = k+1)k\left(y_i|\theta_{k+1}\right)}{p'(s_i = k+1)k\left(y_i|\theta_{k+1}\right) + p'(s_i = j^\star)k\left(y_i|\theta_{j^\star}\right)}.$$

The acceptance probability is

$$\min\left\{1, \frac{k-1}{\lambda(\max(t_i) - \min(\tau_i))} \prod_{\{i|s_i' \neq s_i\}} \frac{k(y_i|\theta_{k+1}')}{k(y_i|\theta_{j^\star})} \prod_{i=1}^{n} \frac{p'(s_i')}{p(s_i)} q^\star\right\}$$

where

$$q^\star = \prod_{\{i|s_i = j^\star\}} \left(\frac{1}{q_{i,1}}\right)^{\mathrm{I}(s_i' = j^\star)} \left(\frac{1}{q_{i,2}}\right)^{\mathrm{I}(s_i' = k+1)}$$

and p' is calculated using the proposal and p is calculated using the current state.

The delete move proposes to remove a point of the process by uniformly selecting two distinct points j_1 from the $\{1, 2, \ldots, k\}/\{i|\tau_i \leq \tau_j \text{ for all } j\}$ and j_2 from $\{1, 2, \ldots, k\}$. We propose to remove τ_{j_1}, θ_{j_1}, and the vectors w_{j_1} and ϵ_{j_1}. For a points $\tau_i < \tau_{j_1}$, we propose new vectors ϵ_i' and w_i' by deleting the element $\epsilon_{i,m}$ where $m = \#\{j|\tau_j > \tau_i\}$ from ϵ_i and the element $w_{i,m}$ where $m = \#\{j|\tau_j > \tau_i\}$ from w_i. Finally, we set $s_i' = j_2$ if $s_i = j_1$. The acceptance probability is zero if $\tau_{j_2} > t_i$ for any i such that $s_i = j_1$. Otherwise, the acceptance probability is

$$\min\left\{1, \frac{\lambda(\max(t_i) - \min(\tau_i))}{k} \prod_{\{i|s_i' \neq s_i\}} \frac{k(y_i|\theta_{j_2})}{k(y_i|\theta_{j_1})} \prod_{i=1}^{n} \frac{p'(s_i')}{p(s_i)} q^\star\right\}$$

where

$$q^\star = \prod_{\{i|s_i = j_1 \text{ or } s_i = j_2\}} \left(\frac{1}{q_{i,1}}\right)^{\mathrm{I}(s_i' = j_1)} \left(\frac{1}{q_{i,2}}\right)^{\mathrm{I}(s_i' = j_2)}.$$

The reverse proposals $q_{i,1}$ and $q_{i,2}$ are calculated as

$$q_{i,1} = \frac{p(s_i = j_1)k\left(y_i|\theta_{j_1}\right)}{p(s_i = j_1)k\left(y_i|\theta_{j_1}\right) + p(s_i = j_2)k\left(y_i|\theta_{j_2}\right)}$$

and

$$q_{i,2} = \frac{p(s_i = j_2)k\left(y_i|\theta_{j_2}\right)}{p(s_i = j_1)k\left(y_i|\theta_{j_1}\right) + p(s_i = j_2)k\left(y_i|\theta_{j_2}\right)}.$$

The Move step uses a Metropolis-Hastings random walk proposal. A distinct point are chosen at random from the set $\{1, 2, \ldots, k\}/\{i|\tau_i \leq \tau_j \text{ for all } j\}$, say $j^\star$, and a new value $\tau'_{j^\star} = \tau_{j^\star} + \epsilon$ where $\epsilon \sim \mathrm{N}(0, \sigma^2_{PROP})$. The move is rejected if $\tau'_j < \min(\tau_j)$, $\tau'_j > \max(t_i)$ or $\tau'_{j^\star} > t_i$ for any i such that $s_i = j^\star$. Otherwise, the acceptance probability is

$$\min\left\{1, \prod_{i=1}^{n} \frac{p'(s'_i)}{p(s_i)}\right\}.$$

Updating λ

The parameter λ can be updated in the following way. Let $\tau^{(old)}$ and $\lambda^{(old)}$ be the current values in the Markov chain. Simulate λ from the distribution proportional to

$$p(\lambda)\lambda^{\#\{i|\tau_i > \min\{t_i\}\}} \exp\{-\lambda(\max(t_i) - \min(t_i)\}$$

and set $\tau_i = \min(t_i) - \frac{\lambda^{(old)}}{\lambda}(\min(t_i) - \tau_i^{(old)})$ if $\tau_i^{(old)} < \min(t_i)$.

Updating θ

The parameter θ_j can update from the full conditional distribution

$$h(\theta_j) \prod_{\{i|s_i=j\}}^{n} k(y_i|\theta_j).$$

B.2 Poisson-Dirichlet process

In this case the general sampler can be simplified to a method that generalizes the computational approach described by Dunson *et al* (2007) for a process where each break is formed by the product of two beta random variables. In our more general case we can write

$$V_j(t) = \epsilon_j \prod_{h=2}^{m_j(t)+1} w_{j,h}.$$

We introduce latent variables r_{ijk} which takes values 0 or 1 where

$$p\left(r_{ij1} = 1\right) = \epsilon_j, \qquad p\left(r_{ijk} = 1\right) = w_{j,k} \text{ for } k = 2, \ldots, m_j(t_i) + 1$$

which are linked to the usual allocation variables s_i by the relationship

$$s_i = \min\{j|r_{ijk} = 1 \text{ for all } 1 \leq k \leq m_j(t_i) + 1\}.$$

Thus

$$\begin{aligned} p(s_i = j) &= p(r_{ijk} = 1 \text{ for all } 1 \leq k \leq m_j(t_i)) \prod_{l \leq j} p(\text{there exists } k \text{ such that } r_{ilk} = 0) \\ &= p(r_{ijk} = 1 \text{ for all } 1 \leq k \leq m_j(t_i)) \prod_{l \leq j} (1 - p(r_{ilk} = 1 \text{ for all } 1 \leq k \leq m_l(t_i))) \\ &= \epsilon_j \prod_{i=2}^{m_j(t_i)} w_{j,i} \left(1 - \epsilon_l \prod_{i=2}^{m_l(t_i)} w_{j,i}\right). \end{aligned}$$

Conditional on r the full conditional distributions of ϵ and w will follow beta distributions and any hyperparameters of the stick-breaking process can be updated using standard technique. Updating of the other parameters proceeds by marginalising over r but conditioning on s.

Updating s

We could update s using the method in Appendix B.1 but we find that this can run very slowly in the Poisson-Dirichlet case. This is because at each update of s_i we potentially simulate a proposed value s_i' which is much bigger than $\max\{s_i\}$ (due to the slow decay of Poisson-Dirichlet processes) and generate very many values for w. This section describes an alternative approach which updates in two steps: 1) update s_i marginalising over any new ϵ and w vectors and 2) simulate the new ϵ and w vectors conditional on the new value of s_i. The algorithm is much more efficient since many proposed values of s_i are rejected at stage 1) and extensive simulation is avoided. We make the following changes to the algorithm

$$q_j = \alpha \frac{1-b}{1+a+b(j-1)} \prod_{k_{-i}<l<j} \frac{a+bl}{1+a+b(l-1)} \prod_{l \leq k_{-i}} (1-V_j(t_i)), \qquad j > k_{-i}$$

and sequentially simulate $(\theta'_{k_{-i}+1}, \tau'_{k_{-i}+1}), (\theta'_{k_{-i}+2}, \tau'_{k_{-i}+2}), \ldots, (\theta'_m, \tau'_m)$ in the same way as before until we meet the condition that $u < \sum_{j=1}^{m} q_j$. The new state (θ', τ', s_i') is

$$\zeta = \begin{cases} 1 & \text{if } \max\{s_i\} \leq k \\ \frac{k(y_i|\theta'_m)}{\alpha} & \text{if } \max\{s_i\} > k \end{cases}.$$

If the move is accepted we simulate ϵ_j and w_j for $j > k_{-i}$ in the following way. Simulate r_{ijk} where $k = \#\{l|\tau_j < \tau_l < t_i\}$ for $k_{-i} < j \leq s_i$ and simulate ϵ_j and w_j using the method for updating ϵ and w.

Updating ϵ and w

We can generate r conditional on s using the following scheme. For the i-th observation, we can simulate $r_{ij1}, r_{ij2}, \ldots, r_{ijk}$ where $k = \#\{l|\tau_j < \tau_l < t_i\}$ sequentially. Initially,

$$p(r_{ij1} = 1) = \epsilon_j \frac{1 - \prod_{h=1}^{k} w_{j,h}}{1 - \epsilon_j \prod_{h=1}^{k} w_{j,h}}.$$

To simulate r_{ijl}, then if $r_{ijh} = 1,\ 1 \leq h \leq l$ then

$$p(r_{ijl} = 1) = w_{j,l} \frac{1 - \prod_{h=l+1}^{k} w_{j,h}}{1 - \prod_{h=l}^{k} w_{j,h}}.$$

Otherwise $p(r_{ijl} = 1) = w_{j,l}$. Finally, we set $r_{i(k+1)1} = 1, \ldots, r_{i(k+1)k} = 1$. Then the full conditional distribution of ϵ_j is

$$\text{Be}\left(1 - b + \sum_{\{i|1<\#\{\tau_k|\tau_j \leq \tau_k < t_i\}\}} r_{ij1}, a + b + \sum_{\{i|1<\#\{\tau_k|\tau_j \leq \tau_k < t_i\}\}} (1 - r_{ij1})\right)$$

and the full conditional distribution of $w_{j,k}$ for $k \geq 1$ is

$$\text{Be}\left(1+a+(k-2)b+\sum_{\{i|k<\#\{\tau_k|\tau_j\leq\tau_k<t_i\}\}} r_{ijk}, b+\sum_{\{i|k<\#\{\tau_k|\tau_j\leq\tau_k<t_i\}\}}(1-r_{ijk})\right).$$

Updating τ, θ and λ

These can be updated using the methods in Appendix B.1.

B.3 Dirichlet process - Marginal method

Updating s

We can update s_j conditional on $s_1,\ldots,s_{j-1},s_{j+1},\ldots,s_n$. We define $A(t)$ to be the active set defined using the allocations $s_1,\ldots,s_{j-1},s_{j+1},\ldots,s_n$, $\mathcal{T}=\{\tau_{s_l}|l=1,2,\ldots,j-1,j+1,\ldots,n\}$, $\mathcal{S}=\mathcal{T}\cup\{t_1,t_2,\ldots,t_n\}$ and $k^\star(y_j)=\int k(y_j|\theta)h(\theta)\,d\theta$. We use the discrete distribution derived from Theorem 3. Let $\tau_1^\star<\tau_2^\star<\cdots<\tau_k^\star$ be the elements of $\mathcal{T}$, where k is the size of $\mathcal{T}$, and let $\phi_1<\phi_2<\cdots<\phi_l$ be the elements of $\mathcal{S}$. We define

$$\begin{aligned}C_i &= \exp\left\{-\frac{\lambda M^2(\phi_{i+1}-\phi_i)}{(M+A(\phi_{i+1}))^2(1+M+A(\phi_{i+1}))}\right\}\\ &= \rho^{\frac{M^2(M+1)}{(M+A(\phi_{i+1}))^2(1+M+A(\phi_{i+1}))}(\phi_{i+1}-\phi_i)}, \qquad 1\leq i<l\end{aligned}$$

and

$$D_i=\frac{M+A(\tau_i^\star)}{1+\eta_i^{-j}+M+A(\tau_i^\star)}, \qquad 1\leq i\leq k$$

where $\eta_i^{-j}=\sum_{k=1,k\neq j}\mathrm{I}(s_k=i)$. Let $\phi_p=t_j$ and $\tau_q^\star$ be the largest element of $\mathcal{T}$ smaller than t_j. The full conditional distribution of s_j is given by

$$p(s_j=m)\propto k(y_j|\theta_m)(1-D_m)\prod_{\{l|\tau_m^\star\leq\phi_l\leq\phi_p\}}C_l\prod_{h=j+1}^{q}D_h, \qquad 1\leq m\leq q.$$

$$p(s_j=k+1 \text{ and } \tau_{k+1}^\star\in(\phi_i,\phi_{i+1}))\propto k^\star(y_j)\,(1-C_i)\prod_{h=i+1}^{p}C_h\prod_{\left\{\tau_h^\star|\phi_i<\tau_h^\star\leq\phi_p\right\}}D_h$$

$$p\left(s_j=k+1 \text{ and } \tau_{k+1}^\star\in(-\infty,\phi_1)\right)\propto k^\star(y_j)\prod_{h=1}^{p}C_h\prod_{i=1}^{q}D_i.$$

If $s_j=k+1$ the new time $\tau_{k+1}^\star$ needs to be drawn in the following way: If $\tau_{k+1}^\star\in(-\infty,\phi_1)$, $\tau_{k+1}^\star=\phi_1-x$ where $x\sim\text{Ex}\left(\frac{\lambda}{M+1}\right)$ and if $\tau_{k+1}^\star\in(\phi_i,\phi_{i+1})$, $\tau_{k+1}^\star=\phi_{i+1}-x$ where $x\sim\text{TEx}_{(0,\phi_{i+1}-\phi_i)}\left(\frac{\lambda M^2}{(M+A_n(\phi_{i+1}))^2(M+A_n(\phi_{i+1})+1)}\right)$ and $\text{TEx}_{(a,b)}(\lambda)$ represent an exponential distribution truncated to (a,b) as defined in Theorem 3.

Updating $\tau^\star$

We can update $\tau_i^\star$ from its full conditional distribution. We define $\phi_1 < \phi_2 < \cdots < \phi_l$ be the elements of $\{t_1, t_2, \ldots, t_n\} \cup \{\tau_1^\star, \ldots, \tau_{i-1}^\star, \tau_{i+1}^\star, \ldots, \tau_k^\star\}$. Let $A(t)$ be the active set excluding the values of j for which $\tau_{s_j} = \tau_i$, K by $\phi_K = \min\{t_k | s_k = i\}$ and $\eta_i = \#\{k | s_k = i\}$. Let

$$P_j = \exp\left\{-\frac{\lambda M(\phi_{j+1} - \phi_j)A_n(\phi_j)}{(M + A_n(\phi_{j+1}))^2}\right\},$$

$$P_j' = \exp\left\{-\frac{\lambda M(\phi_{j+1} - \phi_j)(A_n(\phi_{j+1}) + \eta_i)}{(M + A_n(\phi_{j+1}))(\eta_i + M + A_n(\phi_{j+1}))}\right\}, \qquad 1 \le j < l,$$

$$Q_j = \frac{\Gamma(M + A_n(\tau_j)}{\Gamma(1 + \eta_j + M + A_n(\tau_j))}, \quad Q_j' = \frac{\Gamma(M + A_n(\tau_j) + \eta_i)}{\Gamma(M + A_n(\tau_j) + \eta_i + 1 + \eta_j)}, \qquad 1 \le j \le k.$$

and

$$A_j = \frac{M\Gamma(M + A_n(\phi_{j+1}))}{\Gamma(1 + \eta_i + M + A_n(\phi_{j+1}))} \exp\left\{-\lambda \frac{M(\phi_{j+1} - \phi_j)(A_n(\phi_{j+1} + \eta_i)}{(M + A_n(\phi_{j+1})(M + A_n(\phi_{j+1}) + \eta_i)}\right\}$$

The probability that $\tau_i^\star \in (\phi_j, \phi_{j+1})$ is proportional to

$$A_j \prod_{h=j+1}^{p} \frac{P_h'}{P_h} \prod_{\{i | \phi_j < \tau_i^\star \le \phi_p\}} \frac{Q_i'}{Q_i}, \qquad j \le K$$

and the probability that $\tau_i^\star < \phi_1$

$$\frac{\Gamma(M+1)}{\eta_i \Gamma(\eta_i + M)} \prod_{h=1}^{p} \frac{P_h'}{P_h} \prod_{i=1}^{q} \frac{Q_i'}{Q_i}.$$

This distribution is finite and discrete and draws can be simply simulated. Conditional on the atom being allocated to the region (ϕ_{j-1}, ϕ_j), then the simulated value $\tau_i^\star = \phi_j - x$ where x is distributed $TEx_{(0, \phi_j - \phi_{j-1})}\left(\lambda(\phi_j - \phi_{j-1})\frac{M}{M + A(\phi_j)}\frac{\eta_i}{M + A(\phi_j) + \eta_i}\right)$ and if $\tau_i^\star < \phi_1$ then $\tau_i^\star = \phi_1 - x$ where $x \sim \text{Ex}(\lambda/(M+1))$.

Updating λ and M

To update these parameters from their full conditional distribution we first simulate the number of atoms, c_i, between (ϕ_i, ϕ_{i+1}) from a Poisson distribution with mean $\frac{M\lambda}{M + A_n(\phi_{i+1})}(\phi_{i+1} - \phi_i)$ and $V_1, V_2, \ldots, V_{k_n}$ where $V_i \sim \text{Be}(1 + \eta_i, M + A_n(\tau_i))$.

The full conditional distribution of λ is proportional to

$$p(\lambda)\, \lambda^{k_n + \sum_{i=1}^{k_n} c_i} \exp\left\{-\lambda(\max\{t_i\} - \min\{\tau_i\})\right\}.$$

The full conditional distribution of M is proportional to

$$p(M)\, M^{k_n + \sum_{i=1}^{k_n} c_i} \prod_{i=1}^{k_n} \frac{M + A_n(\tau_i)}{M + A_n(\tau_i) + 1 + \eta_i} \prod_{i=1}^{k_n+1} \left(\frac{M + A_n(\phi_{i+1})}{1 + M + A_n(\phi_{i+1})}\right)^{c_i}.$$